# APRENDIZADO DE MÁQUINA E PRESTAÇÃO DE SERVIÇOS DE ARMAZENAMENTO DE DADOS: MÉTRICAS PARA ANÁLISE E VALIDAÇÃO DE ALGORITMOS PREVISORES

*MACHINE LEARNING AND PROVISION OF DATA STORAGE SERVICES: METRICS FOR ANALYSIS AND VALIDATION OF PREDICTIVE ALGORITHMS*

Carlos Roberto Souza Carmo[1]
Jéssica Rayse de Melo Silva[2]

**RESUMO:**
Utilizando o processo de aprendizado de máquina (*machine learning*), esta pesquisa teve por objetivo inicial analisar como variáveis relacionadas à quantidade média de dias de funcionamento, à idade média dos dispositivos de armazenamento, às taxas de falhas anualizadas, à capacidade de armazenamento, ao fabricante e ao tipo de dispositivo poderiam caracterizar-se como possíveis determinantes da quantidade de falhas ocorridas nos diversos modelos de *Hard Disk Drive* (HD) e *Solid State Disks* (SSD) utilizados em *datacenters* de provedores de serviços de armazenamento de dados em nuvem. Adicionalmente, procurou-se investigar e identificar um conjunto de métricas voltadas para o processo de análise da eficiência das estimativas realizadas com base neste estudo, assim como, da eficiência de algoritmos de previsão em geral. Inicialmente, implementou-se uma rede neural artificial (RNA) aplicada a um conjunto de dados provenientes de 203.168 HD e 2.200 SSD, agrupados em 31 diferentes modelos de fabricação. Após isso, procedeu-se à análise da qualidade das estimativas realizadas com base em RNA, mediante a utilização de um conjunto de métricas pesquisadas para essa finalidade. Apesar de aparentemente apresentar qualidade analítica promissora nas fases de treinamento e teste, a RNA pesquisada mostrou-se ineficiente do ponto de vista preditivo. Por outro lado, foi possível identificar, propor e testar um conjunto métricas voltadas para a análise e validação da eficiência de algoritmos de previsão baseados em aprendizado de máquina.

**PALAVRAS-CHAVE**: falhas; rede neural; métodos quantitativos e computacionais.

***ABSTRACT:***
*Using the machine learning process, the initial objective of this research was to analyze how variables related to the average number of operating days, the average age of storage devices, annualized failure rates, storage capacity, manufacturer and type of device could be characterized as possible determinants of the number of failures that occurred in the different models of Hard Disk Drive (HD) and Solid State Disks (SSD) used in datacenters of cloud*

[1] Doutor em Agronomia pela UNESP (Botucatu) e Mestre em Ciências Contábeis pela PUC-SP. Professor da Faculdade de Ciências Contábeis da Univ. Federal de Uberlândia (FACIC-UFU). e-mail: carlosjj2004@hotmail.com / crscarmo@ufu.br. Orcid: https://orcid.org/0000-0002-3806-9228

[2] Doutora e mestra em Ciências Contábeis pela PPGCC UFU. Professora da Faculdade de Ciências Contábeis da Univ. Federal de Uberlândia (FACIC-UFU). e-mail: jessicar@ufu.br. Orcid: https://orcid.org/0000-0002-5052-094X.

*data storage service providers. Additionally, an attempt was made to investigate and identify a set of metrics aimed at the process of analyzing the efficiency of the estimates made based on this study, as well as the efficiency of forecasting algorithms in general. Initially, an artificial neural network (ANN) was implemented and applied to a set of data from 203,168 HD and 2,200 SSD, grouped into 31 different manufacturing models. After that, we proceeded to analyze the quality of the estimates based on ANN, using a set of metrics researched for this purpose. Despite apparently presenting promising analytical quality in the training and testing phases, the researched ANN proved to be inefficient from a predictive point of view. On the other hand, it was possible to identify, propose and test a set of metrics aimed at analyzing and validating the efficiency of prediction algorithms based on machine learning.*

***Keywords:*** *failures; neural network; quantitative and computational methods.*

## Introdução

Ferramentas de busca, armazenamento em nuvem, compras eletrônicas, entre tantos outros serviços disponibilizados via internet, têm aproximado usuários, consumidores, fornecedores, entre outros, e elevado consideravelmente a demanda por comunicação e capacidade armazenamento da informação (SANTOS; MAZIERO; BONA, 2015).

Diante disso, os serviços de armazenamento em nuvem funcionam ininterruptamente, demandando disponibilidade e capacidade operacionais de forma contínua. Tanto que, em provedores de serviços de nuvem como Aliyun, Amazon, Microsoft Azure, Backblaze, o de tempo de atividade dos usuários normalmente é de 99,9%, no mínimo (LI *et al*., 2021).

Nesse ambiente de prestação de serviços em larga escala são utilizados diversos tipos de mecanismos de armazenamento, com marcas, tecnologias e modelos variados, cuja finalidade é atender diferentes tipos de carga de trabalho, o que leva à produção de riscos relacionados a falhas de funcionamento das mais variadas naturezas (SHEN *et al*., 2018).

Conforme observam Santos, Maziero e Bona (2015, p. 1), "a tecnologia mais popular de armazenamento secundário ainda é o disco magnético (HDD – *Hard Disk Drive*), que alia grande capacidade de armazenamento e baixo custo [...]". Sendo que, "recentemente, os discos de estado sólido (SSD – *Solid State Disks*) elevaram muito o desempenho no acesso ao armazenamento secundário [...]", porém, "[...]seu custo elevado e baixa capacidade inviabilizam a substituição integral dos discos magnéticos por SSD a curto prazo, sobretudo em instalações de maior porte" (MAZIERO; BONA, 2015, p. 1)

Mas, ainda que com baixas taxas de problemas de funcionamento e/ou com certo grau de tolerância a esse tipo de ocorrência, prestadores de serviços de acesso e armazenamento de dados em larga escala têm especial interesse em compreender a natureza de possíveis falhas

nos seus dispositivos de armazenamento devido a dois motivos específicos: economia de custos, pois mesmo com dispositivos de segurança e suporte, falhas de funcionamento demandam a recuperação transitória de dados e informações, o que gera gastos com reparos propriamente ditos; e, prevenção de ocorrências futuras, uma vez que a compreensão das variáveis-críticas desse tipo de problema pode ajudar a identificar tendências de falhas de curto prazo (CHAKRABORTTII; LITZ, 2021).

Logo, parece ser relevante identificar e compreender os possíveis determinantes das falhas de HD e SSD utilizados em provedores de serviços de nuvem, tanto para a prevenção de ocorrências desse tipo de problema no curto prazo, quanto para a identificação de possíveis tendências de ocorrências futuras.

Além disso, ao se preocupar com a previsão de eventos futuros, os processos de estimativa e análise permitem identificar fatores críticos para o sucesso do planejamento operacional e financeiro, o que tem gerado considerável interesse tanto por parte do mercado profissional quanto por parte de acadêmicos, levando à proposição de novos métodos de previsão ao longo dos últimos anos (BOJER; MELDGAARD, 2021).

Ao considerar que no processo de modelagem propriamente dito pode-se utilizar pouca ou, até mesmo, uma única medida da qualidade de ajuste de modelos analíticos, para o processo de previsão torna-se necessária uma quantidade maior de métricas de avaliação (KARUNASINGHA, 2022).

Nesse contexto, esta pesquisa teve por objetivo inicial utilizar processo de aprendizagem de máquina (*machine learning*) baseado em redes neurais artificiais para identificar como variáveis relacionadas a (i) quantidade média de dias de funcionamento, (ii) idade média dos dispositivos de armazenamento, (iii) taxas de falhas anualizadas (AFR – *annualized failure rate*), (iv) capacidade de armazenamento em terabyte (TB), (v) o fabricante e o (vi) tipo de dispositivo (HD ou SSD) podem caracterizar-se como possíveis determinantes da quantidade de falhas ocorridas nos diversos modelos de HD e SSD utilizados nos *datacenters* de provedores de serviços de nuvem. E, na sequência, buscou-se identificar e propor um conjunto de métricas voltadas para a avaliação da eficiência dessa modelagem proposta a partir do processo de *machine learning*.

Nesse sentido, além de problematizar as falhas ocorridas nos dispositivos de armazenamento utilizados em grandes *datacenters* de internet, considerando aquelas 6 categorias de possíveis direcionadores desse tipo de evento tecnológico decorrente da prestação de serviços de armazenamento de dados em grande escala, e ainda, realizar a

pesquisa e a identificação de métricas voltadas para avaliação de erros de estimativas realizadas com base em algoritmos de previsão, inicialmente, buscou-se promover o respectivo embasamento teórico voltado para suporte ao processo de pesquisa propriamente dito, conforme apresentado na segunda seção deste artigo.

A seguir, foi identificada a base de dados necessária à composição da amostra desta pesquisa, além da definição do método de análise analítico-preditiva da ocorrência de falhas de funcionamento de HD e SSD utilizados em *datacenters* de provedores de serviços de nuvem, bem como, o detalhamento do conjunto de métricas avaliadoras do processo de previsão da ocorrência desse tipo de falha, conforme relatado na terceira seção deste artigo.

A quarta seção deste estudo de natureza científica foi destinada à descrição do processo de análise dos dados e à apresentação e discussão dos resultados observados, sem perder de vista a plataforma teórica constituída e já relatada na segunda seção do presente artigo.

A quinta e última seção deste relatório de pesquisa foi destinada às considerações finais acerca de todo o processo de investigação científica.

## 2 Referencial Teórico

Entre os dispositivos de armazenamento secundário, que são aqueles que permitem a manutenção da informação de maneira não-volátil, os HD são considerados a forma mais recorrente de armazenamento *on-line*; contudo, já há algum tempo, os SSD vêm ganhando espaço devido ao seu melhor desempenho relacionado à velocidade de leitura, entre outros atributos (SANTOS; MAZIERO; BONA, 2015).

Enquanto um HD utiliza uma cabeça de leitura de funcionamento mecânico para acessar a mídia armazenada em um disco magnético rotativo, o SSD utiliza memória não volátil como mídia de armazenamento, o que faz com que eles sejam menos frágeis e mais rápidos que os HD (CORNWELL, 2012).

Devido à escalada sem precedentes dos sistemas de armazenamento, a incidência de falhas de dispositivos como os HD e os SSD tem crescido consideravelmente, o que pode produzir consequências desastrosas como perdas de informação de difícil recuperação ou até irrecuperáveis, além da elevação dos custos dos provedores de serviços de nuvem ou *internet data center* (IDC); por isso, esses IDC monitoram constantemente as condições de trabalho desses dispositivos com o uso de sensores de emissão de acústica, contadores, sensores térmicos, entre outros (SHEN *et al*., 2018).

Entretanto, o sucesso e a precisão desse tipo de monitoramento não são muito promissores, principalmente quando o desejável é a abordagem de caráter preventivo, uma vez que a previsão de falhas deve fornecer condições para que os trabalhos de manutenção (*backup* e substituição) sejam executados em tempo hábil (SHEN *et al*., 2018).

Nesse sentido, a compreensão dos possíveis determinantes das falhas em dispositivos de armazenamento é útil de várias maneiras, por exemplo: na elaboração de critérios de escolha de fornecedores e modelos, com vistas à relação desempenho-capacidade-custo; na redução das consequências inerentes à inatividade de um servidor durante o processo de manutenção; no processo de seleção e compra de peças sobressalentes; bem como, para redirecionamento da carga de trabalho entre dispositivos (NARAYANAN *et al*., 2016).

Ao analisar o processo de previsão de falhas de HD, Shen *et al*. (2021) identificam 3 métodos recorrentemente utilizados, ou seja, o método binário, o método das fases e o método do grau de saúde, isso é: o primeiro (binário) classifica os HD em falho ou bom, o que não é considerado correto uma vez que a deterioração de uma unidade de armazenamento é gradual; o segundo método (classificação em fases) divide o processo de deterioração da unidade de armazenamento em até 6 níveis/fases, nas quais, o nível 6 classifica uma unidade como boa e 1 classifica a unidade prestes a falhar em até 72 horas; e, finalmente, o terceiro método (grau de saúde) implica na construção de funções lineares voltadas para descrição do processo de deterioração da unidade de armazenamento devido ao tempo.

Independentemente do método utilizado para compreender e/ou prever os possíveis direcionadores de falhas de HD e SSD, torna-se imprescindível avaliar e conhecer as possíveis variáveis críticas que podem comprometer o processo de modelagem preditiva desse tipo de evento, ou seja: desequilíbrio amostral; diferenças estruturais; diversidade de modelos e fabricantes, assim como, diferenças de carga de trabalho; a ocorrência de novos tipos de falhas; e, o recorte longitudinal utilizado para coleta de dados.

Em qualquer método utilizado para prever falhas de dispositivos de armazenamento nos IDC, o desequilíbrio amostral é muito significativo, uma vez que a quantidade de unidades falhas é expressivamente muito menor que as unidades íntegras (GABER, 2016; CHAKRABORTTII *et al*., 2021; SHEN *et al*., 2021).

Modelos que classificam falhas de funcionamento de maneira uniforme tendem a ser especialmente falhos, pois existem diversos tipos de falhas devido às diferenças estruturais (eletromecânicas e mecânicas) decorrentes da engenharia de um dispositivo de armazenamento (SHEN *et al*., 2018).

Outro fator crítico do processo de modelagem preditiva das falhas ocorridas em HD e SSD está relacionado aos seus diferentes desempenhos, uma vez que existe uma relevante diversidade de modelos e de fabricantes, e ainda, esses dispositivos de armazenamento estão sujeitos à diferentes cargas de trabalho nos IDC (NARAYANAN *et al*., 2016; SHEN *et al*., 2018).

Existe ainda a possibilidade de ocorrências de falhas de caráter inédito, em decorrência da diversidade e da dinâmica de atividade dos ambientes em que os dispositivos de armazenamento são utilizados; sendo que, nesse caso, os dados de falhas dos SSD merecem especial atenção, uma vez que eles ainda são minoria nos IDC (CHAKRABORTTII *et al*., 2021; NARAYANAN *et al*., 2016).

E ainda, levando-se em conta que a deterioração de um dispositivo de armazenamento é gradual, ocorrendo ao longo do tempo, é importante analisar tanto dados de longo prazo quanto de curto prazo (GABER, 2016; LI *et al*., 2021).

Shen *et al*. (2021) alertam para o fato de que os fabricantes de HD geralmente utilizam algoritmos baseados nas chamadas estatísticas *Self-Monitoring Analysis and Reporting Technology* (SMART) para analisar o estado de saúde dos seus dispositivos de armazenamento e prever a ocorrência de falhas eventuais, recomendando o *backup* e a substituição das unidades; contudo, a taxa de precisão na detecção de falhas varia de 3% a 10%, com uma taxa de 0,1% de alarmes falsos, o que dificulta consideravelmente a eficácia das abordagens preventivas.

O fato é que as falhas dos HD e dos SSD utilizados em IDC podem levar a violações contratuais na prestação desse tipo de serviço, além da ocorrência dos chamados "custos de paralizações de *data centers*" (tradução nossa para "*cost of data center outages*") na ordem US$ 9.000 por minuto, em média, podendo chegar até a US$ 17.000 por minuto (LI *et al*., 2021).

Dessa forma, Gaber (2016) observa que a análise de confiabilidade de unidades de HD e SSD, que implica no monitoramento e aprendizagem de padrões antes da ocorrência das falhas propriamente ditas, é o Santo Graal para qualquer empresa de armazenamento de dados e, por isso, esse assunto é altamente explorado tanto na academia quanto no mercado.

Assim, a partir da utilização do processo de *machine learning* baseado em RNA, espera-se que esta pesquisa possa contribuir para o processo de compreensão acerca das possíveis variáveis determinantes da ocorrência de falhas de HD e SSD utilizados em provedores de serviços de nuvem. Para tanto, o processo analítico adotado nesta investigação

científica utilizou os seguintes parâmetros como possíveis variáveis explicativas: quantidade média de dias de funcionamento, cuja finalidade é captar os efeitos longitudinais da variável tempo sobre o funcionamento dos HD e SSD integrantes da amostra desta pesquisa; idade média dos dispositivos de armazenamento, que tem por objetivo captar os efeitos da deterioração gradual das unidades de HD e SSD integrantes da amostra de pesquisa; taxas de falhas anualizadas ou *annualized failure rate* (AFR), destinadas a captar o efeito da diversidade e da dinâmica de atividade do ambiente em que os dispositivos de armazenamento são utilizados, bem como, das falhas já ocorridas anteriormente em, pelo menos, dois anos anteriores (2020 e 2019), além do período corrente em análise (2021); capacidade de armazenamento em terabyte (TB), que pode permitir captar os efeitos da intensidade da carga de trabalho a que as unidades de armazenamento estão expostas; fabricante, que pode indicar os efeitos relacionados às diferenças eletromecânicas e mecânicas decorrentes da engenharia utilizada nos HD e SSD, e ainda, as decorrentes diferenças de desempenho; e, tipo de dispositivo de armazenamento utilizado, se HD ou SSD, cujo objetivo é captar os efeitos de possíveis ocorrências de falhas inéditas, ou próprias de cada tecnologia, e ainda, o efeito da composição utilizada pelo IDC na distribuição do tipo de dispositivo de armazenamento utilizado na prestação dos seus serviços.

Contudo, é importante ressaltar que devido à essa diversidade de possíveis variáveis explicativas da ocorrência de falhas de HD e SSD, e ainda, em decorrência da pequena quantidade de observações disponíveis para este estudo, optou-se pela utilização de técnicas de *machine learning*, no lugar da aplicação de métodos estatísticos que requerem o atendimento de diversos pressupostos. Tal opção se deve ao fato de que a utilização desse tipo de método computacional implica na obtenção de maior capacidade de processamento e, consequentemente, na maior possibilidade de obtenção de uma solução para o problema proposta nesta pesquisa, em detrimento ao cumprimento de pressupostos estatísticos.

Dessa maneira, torna-se necessário concentrar maior atenção ao processo de aferição da qualidade preditiva da modelagem identificada a partir de técnicas de *machine learning*, como é o caso das redes neurais artificiais *perceptron* de múltiplas camadas (RNA MLP). Ou seja, é preciso ir além das análises voltadas para a distância entre valores observados ($y$) e valores previstos ($\hat{y}$), ou seja, os termos de erros ($e = y-\hat{y}$) identificados a partir de amostras distintas, utilizadas separadamente nas fases de treinamento e de teste de uma RNA MLP. Assim, torna-se necessário considerar a totalidade dos erros ocorridos ($e$) em relação à toda a

amostra de pesquisa, de forma conjunta, e ainda, tanto em termos absolutos quanto relativos, assim como, em termos médios e/ou medianos.

Para tanto, esta investigação científica pesquisou e utilizou 11 medidas diferentes de erro: d*etermination coefficient* ou coeficiente de determinação ($R^2$); *mean absolut error* ou erro absoluto médio (*MAE*); *median absolut error* ou erro absoluto mediano (*MdAE*); *mean absolute percentage error* ou erro percentual médio absoluto (*MAPE*); *median absolute percentage error* ou erro percentual absoluto mediano (*MdAP*E); *symmetric mean absolute percentage error* ou erro percentual médio absoluto simétrico (*SMAPE*); *median symmetric absolute percentage error* ou erro percentual absoluto simétrico mediano *(MdSAPE*); *weighted mean absolute percentage error* ou erro percentual médio absoluto ponderado (*WMAPE*); *mean square error* ou erro quadrático médio (*MSE*); *median square error* ou erro quadrático mediano (*MdSE*); e, *root mean square error* ou raiz do erro quadrático médio (*RMSE*).

Muito utilizado na análise de modelos lineares de regressão, o $R^2$ pode ser entendido como a proporção da variável de estudo explicada pelas variáveis integrantes da respectiva modelagem analítica (McKEAN, J. W.; SIEVERS, 1987). Comumente utilizado como uma medida de adequação de modelos de regressão linear, o $R^2$ deve ser empregado com muita cautela, uma vez que seus altos valores ($R^2 \approx 1$) não necessariamente significam um bom ajuste do modelo previsor aos dados observados na realidade, sendo que, o inverso ($R^2 \approx 0$) também é verdadeiro, pois ele pode ser muito influenciado por *outliers* (MARTINS, 2018).

O *MAE* é uma medida de erro absoluto ($|y-\hat{y}|$) que leva em conta a quantidade total de observações/previsões, portanto, é expresso nessa mesma unidade de medida; logo, sua principal limitação é que ele não permite a análise comparativa de modelos com unidades de medida diferentes, e ainda, tende a penalizar mais previsões com erros maiores (HYNDMAN; KOEHLER, 2006; KARUNASINGHA,2022).

Como uma medida de erro alternativa ao *MAE*, o *MdAE* caracteriza-se pela mediana dos erros absolutos ($|y-\hat{y}|$) ocorridos a cada observação/previsão, portanto, sem levar em conta a quantidade total de observações/previsões; ou seja, ele é uma medida de posição (central) (HYNDMAN; KOEHLER, 2006), o que permite que ele seja utilizado comparativamente ao MAE para identificar a possível influência de grandes erros (extremos).

Calculado inicialmente a partir da média de cada observação e o respectivo erro individual ($[|y-\hat{y}|]/|y|$), O *MAPE* caracteriza-se pela média dos erros percentuais absolutos, o que facilita pra se realizar comparações entre modelos preditivos cujas variáveis de interesse

apresentam unidades de medidas diferentes; por outro lado, ele pode apresentar-se muito enviesado quando as observações de Y tenderem a zero (HYNDMAN; KOEHLER, 2006; HUANG *et al*., 2023).

Semelhante à relação entre o *MAE* e o *MdAE*, o *MdAPE* caracteriza-se pela mediana do erro percentual absoluto e, nesse sentido, ambos (*MAPE* e o *MdAPE*) têm a desvantagem de penalizar mais os erros positivos do que em erros negativos (HYNDMAN; KOEHLER, 2006; HUANG *et al*., 2023), apesar de servir de parâmetro de comparação para "calibrar" leituras baseadas no valor do *MAPE*.

O *SMAPE* representa o erro percentual médio absoluto simétrico, uma vez que no seu denominador, ele não de leva em conta os valores das respectivas observações (ou seja, o $y$ no denominador de $[|y-\hat{y}|]/|y|$), ele leva em conta a média entre valores absolutos observados ($|y|$) e valores absolutos previstos ($|\hat{y}|$) (HYNDMAN; KOEHLER, 2006), o que suaviza o reflexo de erros muito elevados e o efeito dos valores negativos e positivos. De forma análoga ao *MdAE* e ao *MdAPE*, o *MdSPE* caracteriza-se pela mediana do *SMAPE*.

O *WMAPE* também é uma medida de erro percentual (ou decimal) médio absoluto, porém, ele é ponderado tanto pela quantidade de observações quanto pelos respectivos montantes, e, por ser relativo, ele é livre de escala, o que ele facilita comparações entre modelos com variáveis de interesse medidas em unidades distintas (BUSARI; LIM, 2021; FORBES, 2023).

O *MSE* e o *RMSE* caracterizam-se pelo erro quadrático médio e por sua raiz quadrada, respetivamente. Se o *MSE* é expresso em uma unidade de leitura de difícil compreensão, a raiz quadrada calculada no *RMSE* faz com que ele seja expresso na mesma unidade de medida das observações analisadas, o que facilita sua interpretação; tanto o *RMSE* quanto *MAE* penalizam erros muito elevados, mas, tendem a se aproximar um do outro quando os erros são menores (DAM *et al*., 2022; KARUNASINGHA, 2022). E, de forma análoga as todos as medidas de erro mediano (*Md*) já expostas (*MdAE*, *MdAPE* e *MdSPE*), o *MdSE* caracteriza-se pela mediana do erro quadrático (*MSE*) (HYNDMAN; KOEHLER, 2006).

Figura 1 – Métricas de precisão utilizadas para avaliar as previsões realizadas com base em RNA

| **Descrição** | **Fórmula** | **Unidade de medida (un) e parâmetro de decisão** |
|---|---|---|
| *Determination coefficient* ou coeficiente de determinação ($R^2$) | $R^2\,(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ | $0 < R^2 < 1$ **ou** $0\% < R^2 < 100\%$ sendo que: quanto mais próximo de 1 melhor |

| | | |
|---|---|---|
| *Mean absolut error* ou erro absoluto médio (*MAE*) | $MAE\ (y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$ | *un do MAE = un de y*<br>sendo que:<br>quanto menor melhor |
| *Median absolut error* ou erro absoluto mediano (*MdAE*) | $MdAE\ (y,\hat{y}) = \lvert y_i - \hat{y}_i\rvert$ | *un do MAE = un de y*<br>sendo que:<br>quanto menor melhor |
| *Mean absolute percentage error* ou erro percentual médio absoluto (*MAPE*) | $MAPE\ (y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{\lvert y_i\rvert}$ | 0% < $MAPE$ < 100%<br>sendo que:<br>quanto menor melhor |
| *Median absolute percentage error* ou erro percentual absoluto mediano (*MdAPE*) | $MdAPE\ (y,\hat{y}) = \frac{\lvert y_i - \hat{y}_i\rvert}{\lvert y_i\rvert}$ | 0% < $MdAPE$ < 100%<br>sendo que:<br>quanto menor melhor |
| *Symmetric mean absolute percentage error* ou erro percentual médio absoluto simétrico (*SMAPE*) | $SMAPE\ (y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{(\lvert y_i\rvert + \lvert\hat{y}_i\rvert)/2}$ | 0% < $SMAPE$ < 100%<br>sendo que:<br>quanto menor melhor |
| *Median symmetric absolute percentage error* ou erro percentual absoluto simétrico mediano (*MdSAPE*) | $MdSAPE\ (y,\hat{y}) = \frac{\lvert y_i - \hat{y}_i\rvert}{(\lvert y_i\rvert + \lvert\hat{y}_i\rvert)/2}$ | 0% < $MdSAPE$ < 100%<br>sendo que:<br>quanto menor melhor |
| *Weighted mean absolute percentage error* ou erro percentual médio absoluto ponderado (*WMAPE*) | $WMAPE\ (y,\hat{y}) = \frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{\sum_{i=1}^{n}\lvert y_i\rvert}$ | 0% < $WMAPE$ < 100%<br>sendo que:<br>quanto menor melhor |
| Mean square error ou erro quadrático médio (*MSE*) | $MSE\ (y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ | *un do MSE = ( y – ŷ )²*<br>sendo que:<br>quanto menor melhor |
| *Median square error* ou erro quadrático mediano (*MdSE*) | $MdSE\ (y,\hat{y}) = (y_i - \hat{y}_i)^2$ | *un do MdSE = ( y – ŷ )²*<br>sendo que:<br>quanto menor melhor |
| *Root mean square error* ou raiz do erro quadrático médio (*RMSE*) | $RMSE\ (y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ | *un do RMSE = un de y*<br>sendo que:<br>quanto menor melhor |
| **Legenda**:<br>*y* = valor da variável analisada;<br>$\hat{y}$ = valor previsto para a variável analisada com base em um modelo;<br>*n* = quantidade total de observações referentes a *y* e/ou $\hat{y}$ ; e<br>*i* = cada observação específica de *y* e/ou $\hat{y}$ . | | |

**Fonte**: elaborado pelos autores com base em Mckean e Sievers (1987), Hyndman e Koehler (2006), Bruce e Bruce (2019), Martins (2018), Harrison (2020), Bojer e Meldgaard (2021) e Karunasingha (2022).

Assim, o conjunto das métricas de precisão pesquisadas e utilizadas para avaliar as previsões realizadas com base em RNA propostas segundo a presente investigação, e ainda, suas unidades de medida e os respectivos parâmetros de decisão foram resumidos no quadro descrito pela Figura 1.

## 3 Metodologia de Pesquisa

A Backblaze é uma empresa americana que atua no segmento de armazenamento em nuvem, *backup* de negócios e *backup* pessoal, e, desde 2013, publica estatísticas e *insights* com base nos dados dos dispositivos de armazenamento do seu IDC, além de disponibilizar

publicamente os dados que subsidiam as informações contidas em seus relatórios, de tal forma que qualquer um possa utilizá-los livremente (BACKBLAZE, 2022a).

Em 2021, a Backblaze (2022b) adicionou 40.460 HD ao seu IDC, totalizando 206.928 unidades a partir de dezembro de 2021, das quais 203.168 foram destinadas exclusivamente ao armazenamento de dados, e ainda, tiveram suas estatísticas disponibilizadas no relatório "*Backblaze Drive Stats for 2021*". Sendo que, também em 2021, a Blackblaze (2022c) compartilhou os dados dos 2.200 SSD utilizados em seu IDC, a partir do seu relatório "*The SSD edition: 2021 drive stats review*".

Dessa maneira, todos os dados integrantes da amostra desta pesquisa tiveram como base as informações disponibilizadas pela Backblaze (2022a; 2022b; 2022c) nos seus relatórios anuais de 2021, conforme detalhado na Tabela 1.

**Tabela 1 –** Dados da amostra de pesquisa

| Qtd. de falhas (un) | Dias de funcion. (un) | Idade média (meses) | AFR 2021 (%) | AFR 2020 (%) | AFR 2019 (%) | TB (TB) | HGST (bin.) | Seagate (bin.) | Toshiba (bin.) | WDC (bin.) | Crucial (bin.) | Micron (bin.) | HD (bin.) | SSD (bin.) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 190 | 13944421 | 66,92 | 0,58 | 0,27 | 0,59 | 4,00 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 309 | 28880152 | 62,37 | 0,31 | 0,27 | 0,44 | 4,00 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 25 | 1607847 | 44,85 | 1,80 | 1,41 | 2,00 | 4,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 20 | 2096141 | 27,04 | 2,04 | 2,01 | 0,00 | 4,00 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 34 | 3581368 | 9,40 | 0,11 | 0,23 | 0,96 | 6,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 140 | 10839720 | 32,95 | 0,64 | 0,29 | 0,79 | 8,00 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 4461 | 66434125 | 74,37 | 1,46 | 0,93 | 1,26 | 8,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 86 | 3549496 | 80,85 | 1,49 | 1,22 | 1,56 | 8,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 574 | 18943397 | 62,63 | 2,26 | 1,33 | 1,00 | 10,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 839 | 23710079 | 52,82 | 0,27 | 0,31 | 0,56 | 12,00 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 68 | 1882831 | 50,07 | 0,29 | 1,19 | 0,00 | 12,00 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1953 | 35588859 | 25,80 | 0,48 | 0,46 | 0,40 | 12,00 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 379 | 12987266 | 21,13 | 2,01 | 1,04 | 3,31 | 12,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 87 | 5066595 | 13,84 | 1,08 | 1,01 | 1,14 | 12,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 98 | 3508278 | 11,10 | 0,52 | 0,84 | 0,00 | 12,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 78 | 592111 | 12,86 | 1,03 | 1,04 | 0,00 | 14,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 49 | 1547263 | 7,74 | 4,79 | 0,00 | 0,00 | 14,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 7 | 306763 | 79,32 | 0,77 | 0,91 | 0,65 | 14,00 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 362 | 16396273 | 14,28 | 1,66 | 0,00 | 0,00 | 14,00 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 7 | 156221 | 11,81 | 0,43 | 0,16 | 0,00 | 14,00 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 4 | 237692 | 3,57 | 1,11 | 1,71 | 0,00 | 16,00 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 12 | 607500 | 8,52 | 0,91 | 0,00 | 0,00 | 16,00 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 36 | 3184727 | 12,81 | 0,70 | 0,00 | 0,00 | 16,00 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 1 | 266410 | 5,06 | 0,14 | 0,00 | 0,00 | 16,00 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 2 | 1689 | 0,40 | 43,22 | 0,00 | 0,00 | 0,50 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 7 | 33478 | 14,00 | 7,63 | 0,00 | 0,00 | 0,24 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 8 | 276281 | 11,10 | 1,06 | 0,00 | 0,00 | 0,25 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1267 | 33,00 | 28,81 | 0,00 | 0,00 | 2,00 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 204287 | 21,70 | 0,36 | 1,18 | 0,00 | 0,30 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 6515 | 36,70 | 0,00 | 0,00 | 5,75 | 0,50 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 39147 | 28,70 | 0,93 | 0,93 | 0,00 | 0,30 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

Fonte: elaborado pelos autores a partir dos dados disponíveis em Blackblaze (2022a; 2022b; 2022c).

A variável “qtd. de falhas”, que é o objeto inicial deste estudo, diz respeito à quantidade total de falhas relatadas pela Backblaze (2022b) para cada modelo de HD utilizado ao longo do período compreendido entre 20/04/2013 a 31/12/2021, bem como, a quantidade total de falhas ocorridas em cada modelo dos seus SSD desde o início da sua utilização em 2021, até 31/12/2021 (BACKBLAZE, 2022c). Dessa forma, essa variável apresenta o total referente às falhas ocorridas ao longo de todo o período de funcionamento dos respectivos dispositivos de armazenamento.

As variáveis “dias de funcionamento“e “idade média” referem-se ao tempo de atividade de cada modelo de dispositivo (média em dias) e a respectiva idade (média em meses), respectivamente (BACKBLAZE, 2022b; 2022c).

As variáveis “AFR 2021”, “AFR 2020” e “AFR 2019” referem-se às taxas de falhas anualizadas dos anos 2021, 2020 e 2019 (percentual médio), respectivamente, calculadas pela Blackblaze (2022b) conforme descrito na Equação 1.

$$\text{AFR} = (\text{falhas na unidade} / (\text{dias de condução} / 365)) \, 100 \qquad (1)$$

A variável “TB” refere-se à capacidade total de armazenamento (média em TB) dos dispositivos utilizados pela Backblaze (2022b; 2022c) no seu IDC.

As variáveis “HGST”, “Seagate”, “Toshiba”, “WDC”, “Crucial” e “Micron” são variáveis binárias, para as quais 1 (um) indica que o modelo pertence a determinado fabricante, e 0 (zero) em caso contrário (BACKBLAZE, 2022b; 2022c).

As variáveis “HD” e “SSD” também são variáveis binárias, para as quais, 1 (um) indica se o modelo em questão é um HDD (*Hard Disk Drive*) ou um SSD (*Solid State Disks*) e 0 (zero) em caso contrário, dependendo da respectiva situação (BACKBLAZE, 2022b; 2022c).

Ao considerar a quantidade de observações disponíveis para análise ($n$ = 31 observações), optou-se por utilizar técnicas de *machine learning* baseadas em redes neurais artificiais (RNA) do tipo *perceptron* multicamada, em detrimento de modelos estatísticos convencionais cujos parâmetros necessários para sua validação tendem a não favorecer pequenas amostras, com especial atenção aos graus de liberdades necessários para tanto.

O processo de *machine learning* baseado em RNA pode ser utilizado para a solução de problemas combinatórios de naturezas diversas, bem como, para a execução de tarefas muito

complexas como o processamento de informações e reconhecimento de padrões, entre outras possibilidades (BRAGA; CARVALHO; LUDERMIR, 2014).

Uma RNA *perceptron* de múltiplas camadas (RNA MLP) "[…] possui uma estrutura composta basicamente por três camadas ou mais camadas, sendo, respectivamente, uma camada de entrada, uma ou mais camadas intermediárias ou ocultas e uma camada de saída" (CARNEIRO JÚNIOR; SOUZA, 2019, p. 221). E ainda, "na camada de entrada são inseridos os parâmetros das variáveis que serão previsoras no processo, ou seja, as variáveis independentes, na camada intermediária ocorrem o processamento dos neurônios e ajustes das funções e pesos sinápticos e, na camada de saída, estão os parâmetros a serem previstos […]" (CARNEIRO JÚNIOR; SOUZA, 2019, p. 221).

Nesse sentido, todos os parâmetros utilizados para a construção da RNA MLP implementada nesta investigação e a descrição das parcelas da amostra de pesquisa escolhidas aleatoriamente para treinamento (aprendizagem) e validação (teste) estão descritas na Figura 2.

**Figura 2 -** Parâmetros da rede neural artificial (RNA) e seu processamento

| | | | | |
|---|---|---|---|---|
| **Camada de entrada** | Fatores | 1 | HGST | |
| | | 2 | Seagate | |
| | | 3 | Toshiba | |
| | | 4 | WDC | |
| | | 5 | Crucial | |
| | | 6 | Micron | |
| | | 7 | HD | |
| | | 8 | SSD | |
| | Covariáveis | 1 | Dias_de_func | |
| | | 2 | Idade_média | |
| | | 3 | AFR_2021 | |
| | | 4 | AFR_2020 | |
| | | 5 | AFR_2019 | |
| | | 6 | TB | |
| | Número de unidades sem a unidade de viés | | 22 | |
| | Método de reescalonamento para covariáveis | | Padronizado | |
| **Camadas ocultas** | Número de camadas ocultas | | 1 | |
| | Número de Unidades na Camada Oculta 1, sem a unidade de viés | | 9 | |
| | Função de ativação | | Tangente hiperbólica | |
| **Camada de saída** | Variáveis dependentes | 1 | Qtd_falhas | |
| | Número de unidades | | 1 | |
| | Método de reescalonamento para dependentes de escala | | Padronizado | |
| | Função de ativação | | Identidade | |
| | Função de erro | | Soma dos Quadrados | |
| **Resumo do processamento** | Amostra | Detalhes | n | Porcentagem |
| | | Treinamento | 23 | 74,20% |
| | | Testes | 8 | 25,80% |
| | Válido | | 31 | 100,00% |
| | Excluídos | | 0 | 0,00% |

| | | | | |
|---|---|---|---|---|
| | Total | | 31 | 100,00% |

**Fonte**: elaborado pelos autores, com base nos dados da pesquisa.

Para a implementação da RNA MLP utilizou-se um computador com processador Intel ® Core™ i3-1005 G1, CPU @ 1.20 GHz e 1.19 GHz, com 4,00 GB de memória RAM instalada, cujo custo de processamento (treinamento e teste) foi inferior a um minuto.

As métricas utilizadas para avaliar a precisão das previsões realizadas com base na RNA MLP proposta por esta investigação, suas unidades de medida e os respectivos parâmetros de decisão já foram detalhados no quadro descrito anteriormente pela Figura 1, apresenta no referencial teórico desta pesquisa.

Assim, ao considerar o seu objeto de estudo, a forma como a respectiva amostra foi composta, o método analítico utilizado e suas métricas, esta investigação pode ser considerada uma pesquisa científica de natureza empírico-analítica cuja amostra foi constituída a partir da conveniência e disponibilidade de informações, baseada em métodos computacionais aplicados.

## 4 Análise dos Dados, Apresentação e Discussão dos Resultados

Ao problematizar a quantidade de falhas ocorridas nos HD e SSD utilizados em provedores de serviços de nuvem, tomando como base o respectivo total de ocorrências por modelo, acredita-se que existe maior potencial para a compreensão acerca dos determinantes dos efeitos longitudinais da variável tempo sobre o funcionamento desses dispositivos de armazenamento de dados. Pois os estudos que tratam somente da possibilidade de ocorrência dessas falhas (com variável de estudo binária do tipo “0” ou “1”) tendem a deixar de atribuir a devida importância à variável tempo, priorizando muito mais fatores de natureza técnica.

Além disso, vislumbra-se a possibilidade deste estudo capturar os possíveis efeitos da deterioração gradual dessas unidades de armazenamento, os efeitos da diversidade e da dinâmica de atividade do ambiente em que tais dispositivos são utilizados, eventuais falhas decorrentes da intensidade da carga de trabalho a que as unidades de armazenamento estão expostas, como as diferenças eletromecânicas e mecânicas decorrentes da engenharia utilizada nos HD e SSD podem influenciar a ocorrência das falhas, e ainda, se existe relacionamento entre as ocorrências de falhas e o tipo de dispositivo propriamente dito (HD ou SSD).

Nesse sentido, após utilizar 23 observações para realizar o treinamento da RNA MLP, o que equivale à 74,20% do total da amostra de dados disponível para estudo, e outras 8 observações para testes do modelo identificado, o que equivale a 25,50% da amostra de

pesquisa, observou-se um coeficiente de determinação ($R^2$) de 0,959 referente às quantidades de falhas previstas pela rede neural, comparativamente às observações reais pertencentes à amostra.

Assim, se fosse admitida uma combinação linear das variáveis explicativas, cerca de 96% das observações reais ($R^2$ .100≈95,90%) poderiam ser explicadas pelas previsões dos totais de falhas realizadas pela RNA MLP com base naquelas 8 variáveis qualitativas e 6 variáveis quantitativas.

A princípio, poder-se-ia considerar que um $R^2$ de 0,959 é um forte indício de que o conjunto das variáveis explicativas utilizadas para compor a modelagem explicativa implementada neste estudo apresenta um bom ajuste linear. E, assim sendo, as informações resumidas na Tabela 2 permitem analisar o grau de importância de cada uma das variáveis independentes consideradas no processo de modelagem da RNA MLP analítico-previsora da quantidade total de falhas ocorridas em cada um dos 31 modelos de HD e SSD integrantes da amostra desta pesquisa.

**Tabela 2 -** Importância das variáveis independentes

| Variáveis | Importância | Importância normalizada |
|---|---|---|
| Dias_de_func | 0,4818 | 100,00% |
| AFR_2021 | 0,0805 | 16,71% |
| AFR_2019 | 0,0683 | 14,17% |
| HGST | 0,0612 | 12,71% |
| Idade_média | 0,0598 | 12,42% |
| HD | 0,0509 | 10,57% |
| TB | 0,0495 | 10,27% |
| WDC | 0,0249 | 5,16% |
| Crucial | 0,0227 | 4,72% |
| SSD | 0,0223 | 4,63% |
| AFR_2020 | 0,0206 | 4,27% |
| Seagate | 0,0204 | 4,24% |
| Micron | 0,0194 | 4,03% |
| Toshiba | 0,0177 | 3,67% |

**Fonte**: elaborado pelos autores, com base nos dados da pesquisa.

Conforme pode ser observado na Tabela 2, a quantidade de dias funcionamento dos HD e SSD foi considerada a variável mais importante no processo explicativo das falhas ocorridas nos dispositivos de armazenamento de dados integrantes da amostra de pesquisa, segundo a RNA MLP implementada neste processo de investigação científica.

Essa evidência pode ser utilizada no suporte à tomada de decisões preventivas relacionadas ao redirecionamento da carga de trabalho realizada por aqueles dispositivos com maior tempo de atividade, conforme sugerido por Narayanan *et al*. (2016).

Adicionalmente, as taxas de falhas anualizadas (AFR) referentes aos anos 2021 e 2019 (percentual médio/modelo) ocuparam o segundo e terceiro lugar em relação ao grau de importância das variáveis explicativo-previsoras integrantes da RNA MLP, respectivamente. Mas, ainda assim, essas duas variáveis não têm o mesmo peso que a variável referente à quantidade de dias de funcionamento dos dispositivos de armazenamento, no processo de previsão das ocorrências de falhas de HD e SSD utilizados em IDC.

Por outro lado, essa evidência vai de encontro ao que foi sugerido por Gaber (2016) e Li *et al*. (2021), ou seja, o monitoramento das taxas de falhas anualizadas (AFR) de períodos passados pode capturar os efeitos da deterioração gradual dos dispositivos de armazenamento. Sendo que, a capacidade preditiva dessa variável pode ser potencializada a partir do monitoramento conjunto com a idade dos respectivos HD e SSD, segundo o de detalhamento do grau de importância das variáveis independentes descrito na Tabela 2.

Ainda segundo as informações resumidas na Tabela 2, é possível perceber que as variáveis representativas dos diferentes fabricantes dos modelos de HD e SSD não são tão expressivas no processo de previsão das ocorrências das falhas, o que também acontece com o fato dos dispositivos utilizados serem HD ou SSD.

Dessa forma, a quantidade de ocorrências de falhas não pode ser considerada um bom critério ou, pelo menos, o único critério para seleção de fabricantes e/ou fornecedores dos dispositivos de armazenamento, conforme sugerido por Narayanan *et al*. (2016).

Cabe observar que, ao utilizar a quantidade total de falhas ocorridas como objeto de estudo, esta pesquisa não enfrentou problemas relacionados ao desequilíbrio amostral decorrente da elevada quantidade de unidades íntegras combinada com a baixa quantidade de unidades de armazenamento falhas, conforme sugerido por Gaber (2016), Chakraborttii *et al*. (2021) e Shen *et al*. (2021).

Ao identificar e classificar as variáveis explicativas de acordo com o respectivo grau de importância no processo analítico das falhas ocorridas nos dispositivos de armazenamento, a RNA MLP implementada nesta pesquisa científica forneceu evidências que podem ser utilizadas em abordagens de caráter preventivo. Como exemplo pode-se citar a possibilidade de redirecionamento da carga de trabalho daquelas unidades de armazenamento com maior quantidade de dias de funcionamento, ou ainda, a parada programada para manutenção e/ou

substituição de HD e SSD que vêm apresentando falhas em anos anteriores, entre outras possibilidades.

Por outro lado, a despeito das evidências coletadas ao longo do processo de modelagem propriamente dito, a RNA MLP pesquisada mostrou-se ineficiente do ponto de vista preditivo, conforme demonstram as métricas descritas na Tabela 3.

**Tabela 2 –** Avaliação da qualidade das previsões realizadas pela RNA MLP, a partir das métricas propostas nesta pesquisa

| Métricas | Valores |
|---|---|
| $R^2$ | 0,959 |
| *MAE (Mean absolut error)* | 86,387 |
| *MdAE (Median absolut error)* | 38,000 |
| *MAPE (Mean absolute percentage error)* | 9,225 |
| *MdAPE (Median absolute percentage error)* | 0,747 |
| *SMAPE (Symmetric mean absolute percentage error)* | 0,967 |
| *MdSAPE (Median symmetric absolute percentage error)* | 1,059 |
| *WMAPE (Weighted mean absolute percentage error)* | 0,272 |
| *MSE (Mean square error)* | 29179,419 |
| *MdSE (Median square error)* | 1444,000 |
| *RMSE (Root mean square error)* | 170,820 |

**Fonte**: elaborado pelos autores, com base nos dados da pesquisa.

Enquanto únicas medidas dos valores absolutos dos erros de estimativas propriamente ditos, o desejável era que *MAE* e *MdAE* se aproximassem. Contudo, segundo os dados analisados nesta investigação, o valor do *MAE* é mais que o dobro do valor do *MdAE*, o que pode ser entendido como um indício de que existem erros com valores bem elevados, o que aumenta consideravelmente a amplitude das respectivas observações. Ou seja, com diferenças entre falhas observadas (reais) e falhas previstas com um valor de erro mínimo de 4 e um valor de erro máximo de 713, perfazendo uma amplitude total de 709, pode-se concluir que tanto o *MAE* quanto o *MdAE* não seriam boas métricas para medir a qualidade do modelo baseado em RNA MLP para previsão de falhas, pois ambos pareceram ter sido influenciados pela magnitude dos erros observados.

Ao realizar a análise das medidas relativas de erro previstas para este estudo, observa-se que também não há muita harmonia entre elas, isso é:

a)o *MAPE* de 9,225 indica um erro percentual médio absoluto na ordem de 922,5%, o que sinaliza que as previsões de falhas realizadas pela RNA MLP apresentaram erros com uma magnitude média 9,22 vezes maiores que as respectivas observações reais;

b)o valor do *MdAPE* é diferente do valor do *MAPE*, corroborando o que aconteceu entre o *MAE* e o *MdAE*, contudo, o valor de 0,747 (74,70% de erro) do *MdAPE* é consideravelmente menor que os 9,225 (ou 922,5%) do *MAPE*, o que reforça ainda mais a evidência da influência daquela elevada amplitude entre os extremos observados para essas medidas de erro;

c)o *SMAPE* de 0,967, enquanto medida erro percentual médio absoluto simétrico, tende a suavizar os efeitos dos valores extremos capturados tanto pelo *MdAPE* quanto pelo *MAPE*, o que faz com ele esteja mais próximo do valor relativo mediano identificado a partir do *MdAPE* e bem mais distante do valor relativo identificado pelo *MAPE*;

d)o MdSAPE de 1,059 reforça a constatação realizada com base no *SMAPE*, pois, apesar de serem diferentes, ainda assim, esses dois valores situam-se relativamente próximos; e

e)apesar de ser o menor valor observado para as medidas relativas de erro, portanto o que mais se distanciou das demais medidas relativas, o *WMAPE* de 0,272 ou 27,20% pôde ser confirmado a partir da divisão do *MAE* pela média das observações reais (*MAE/ȳ*)

Entretanto, o traço comum entre as medidas relativas de erro é que todas capturaram os efeitos dos erros de previsão da RNA MLP proposta, sendo que, pelo menos 2 delas apresentaram certo grau de convergência; ou seja, *SMAPE* (0,967) e *MdSAPE* (1,059) se aproximaram consideravelmente de 1,00, o que sinaliza um erro percentual relativo de 100% entre as previsões realizadas pela RNA MLP e as respectivas observações reais referentes às falhas de dispositivos de armazenamento em análise nesta pesquisa.

Em relação às métricas de erro de previsão baseadas em erro quadrático, o *MSE* (29179,419) e o respectivo valor mediado *MdSE* (1444,000) também se apresentaram bem distantes, ainda que a respectiva unidade de medida não seja muito prática para realizar inferências comparativas com a unidade de medida da variável em estudo. Por outro lado, o *RMSE*, que apresenta unidade medida igual à da variável de estudo, apresentou um erro de 170,820. Esse último valor pode ser comparado com os valores do *MAE* e *MdAE*, sendo que, ele se encontra bem mais próximo daquele primeiro do que do segundo.

De uma maneira geral, o que se percebe é que o modelo baseado em RNA MLP desta investigação não pôde ser considerado um bom previsor da quantidade das falhas ocorridas nos HD e SSD utilizados em provedores de serviços de nuvem ou *internet data center* (IDC).

Em relação às métricas utilizadas para medir essa qualidade, todas foram capazes de sinalizar problemas inerentes às previsões realizadas pela RNA MLP proposta; contudo, dada à diversidade de unidades de medidas, o que ficou evidente foi a necessidade de se utilizar todas de maneira conjunta. Ressaltando-se uma vez mais que, a despeito das respectivas unidades de medidas, houve convergência em relação à ineficiência preditiva da modelagem proposta nesta investigação.

## 5 Considerações Finais

Ao problematizar a quantidade de falhas ocorridas nos diversos modelos de HD e SSD utilizados nos IDC e seus possíveis determinantes, esta investigação científica fez uso de técnicas de *machine learning* baseadas em RNA MLP aplicadas em um conjunto de dados reais provenientes de mais de 208,000 unidades de armazenamento de dados (206.928 HD e 2.200 SSD) agrupadas em 31 diferentes modelos de fabricação.

Ao apresentar uma qualidade analítica relativamente promissora, a RNA MLP implementada nesta investigação permitiu avaliar o grau de importância de cada uma daquelas possíveis variáveis explicativas avaliadas neste estudo de científico de natureza empírico-analítica baseado em métodos computacionais aplicados. Sendo que, ao levar em conta a teoria adjacente, a análise dos resultados permitiu corroborar em parte com os resultados de estudos já realizados anteriormente, e ainda, trouxe algumas evidências que podem contribuir significativamente com o processo de planejamento e tomada de decisões na prestação de serviços de armazenamento de dados em IDC.

Por outro lado, a análise preditiva da modelagem implementada nesta pesquisa não se mostrou capaz de prever satisfatoriamente a quantidade de falhas de dispositivos de armazenamento utilizados em provedores de serviços de nuvem ou IDC. Porém, considerando-se que só foi testada uma modelagem preditiva, não é possível inferir se as divergências (distâncias de valores) entre as métricas utilizadas se devem às observações ou à cada métrica em si.

Como principal limitação desta investigação científica destaca-se o fato da sua amostra ter sido constituída a partir da conveniência relacionada à disponibilidade de informações, o que não permite generalizações acerca dos seus resultados, exceto pela proposição de um conjunto de métricas voltadas para a análise de precisão das previsões realizadas com base em RNA.

Assim, dada a abrangência relacionada à diversidade de modelos de HD e SSD cujas falhas foram analisadas, e ainda, a expressiva quantidade de unidades de armazenamentos cujos dados serviram de base para esta pesquisa, espera-se que as evidências coletadas e analisadas possam ser somadas aos achados de estudos de natureza correlata e, assim, sirvam de contribuição para o debate e a orientação de ações voltadas para o suporte à tomada de decisões na prestação de serviços de tecnologia em grande escala e para a avaliação da qualidade das estimativas realizadas com base em algoritmos de previsão.

## Referências

BACKBLAZE. **Welcome to the Backblaze hard drive data and stats**. San Mateo-CA (EUA), 2022a. Disponível em: https://www.backblaze.com/b2/hard-drive-test-data.html. Acesso em: 08 abr. 2022.

BACKBLAZE. **Backblaze drive stats for 2021**. By Andy Klein, San Mateo-CA (EUA), February 1, 2022b. Disponível em: https://www.backblaze.com/blog/backblaze-drive-stats-for-2021/. Acesso em: 08 abr.2022.

BACKBLAZE. **The SSD edition**: 2021 drive stats review. By Andy Klein, San Mateo-CA (EUA), March 3, 2022c. Disponível em: https://www.backblaze.com/blog/ssd-edition-2021-drive-stats-review/. Acesso em: 08 abr.2022.

BOJER, C. S.; MELDGAARD, J. P.. Kaggle forecasting competitions: an overlooked learning opportunity. **International Journal of Forecasting**, [*s. l.*], v. 37, issue 2, p. 587-603, Apr.-Jun. 2021. Disponível em: https://doi.org/10.1016/j.ijforecast.2020.07.007. Acesso em: 29 nov. 2022.

BRAGA, A. P.; CARVALHO, A. P. L. F.; LUDERMIR, T. B.. **Redes neurais artificiais**: teoria e aplicações. 2ed. Rio de Janeiro: LTC, 2014.

BRUCE, P.; BRUCE, A.. **Estatística prática para cientistas de dados**: 50 conceitos essenciais. Rio de Janeiro-RJ: Alta Books, 2019.

BUSARI, G. A.; LIM, D. H.. Crude oil price prediction: a comparison between AdaBoost-LSTM and AdaBoost-GRU for improving forecasting performance. **Computers & Chemical Engineering**, [*s. l.*], v. 155, e-article 107513, December 2021. Disponível em: https://doi.org/10.1016/j.compchemeng.2021.107513. Acesso em: 27 dez. 2022.

CARNEIRO JÚNIOR, J. B. A.; SOUZA, C. S. de. Aplicação de redes neurais artificiais na previsão do produto interno bruto do Mato Grosso do Sul em função da produção de cana-de-açúcar, açúcar e etanol. **Revista Ibero-Americana de Ciências Ambientais**, [*s. l.*], v. 10, n. 5, p. 218-230, ago.-set. 2019. Disponível em: https://doi.org/10.6008/CBPC2179-6858.2019.005.0019. Acesso em: 08 ago. 2022.

CHAKRABORTTII, C.; LITZ, H.. Explaining SSD failures using anomaly detection. *In*: ANNUAL NON-VOLATILE MEMORIES WORKSHOP (*Online Event*), 12., 2021, San Diego-CA. **Proceedings [...]**. San Diego-CA: University of California/ IEEE Computer Society, mar. 2021. Disponível em: http://nvmw.ucsd.edu/nvmw2021-program/nvmw2021-data/nvmw2021-paper44-final_version_your_extended_abstract.pdf. Acesso em: 24 ago. 2022.

CORNWELL, M.. Anatomy of a solid-state drive. **Communications of the ACM**, [*s. l.*], v. 55, n. 12, p. 59–63, Dec. 2012. Disponível em: https://dl.acm.org/doi/fullHtml/10.1145/2380656.2380672. Acesso em: 23 ago. 2022.

DAM, R. S. de F.; SALGADO, W. L.; SCHIRRU, R.; SALGADO, C. M.. Application of radioactive particle tracking and an artificial neural network to calculating the flow rate in a two-phase (oil–water) stratified flow regime. **Applied Radiation and Isotopes**, [s. l.], v. 180, e-article 110061, February 2022. Disponível em: https://doi.org/10.1016/j.apradiso.2021.11006. Acesso em: 23 dez. 2022.

FORBES, K. F.. Demand for grid-supplied electricity in the presence of distributed solar energy resources: Evidence from New York City. **Utilities Policy**, [*s. l.*], v. 80, e-article 101447, February 2023. Disponível em: https://doi.org/10.1016/j.jup.2022.101447. Acesso em 27 dez. 2022.

GABER. S.. The data science of predicting disk drive failures. **DELL Technologies Services:** Blog, [*s. l.*], e-article, June 27, 2016. Disponível em: https://www.dell.com/en-us/blog/data-science-predicting-disk-drive-failures/. Acesso em: 23 ago. 2022.

HARRISON, M.. **Machine learning**: guia de referência rápida (trabalhando com dados estruturado em Python). São Paulo-SP: Novatec, 2020.

HYNDMAN, R. J.; KOEHLER, A. B.. Another look at measures of forecast accuracy. **International Journal of Forecasting**, [*s. l.*], n. 22, issue 4, p. 679-688, Oct.–Dec. 2006. Disponível em: https://doi.org/10.1016/j.ijforecast.2006.03.001. Acesso em: 29 nov. 2022.

HUANG, S. H. Z.; HIRAOKA, T. ; CHÁVEZ, A. P. de L.; ALA-NISSILA, T; LESKELÄ, L.; KIVELÄ, M.; SARAMÄKI, J..Estimating inter-regional mobility during disruption: Comparing and combining different data sources. **Travel Behaviour and Society**, [*s. l.*], v. 31, p. 93-105, April 2023. Disponível em: https://doi.org/10.1016/j.tbs.2022.11.005. Acesso em: 27 dez. 2022.

KARUNASINGHA, D. S. K.. Root mean square error or mean absolute error? Use their ratio as well. **Information Sciences**, [*s. l.*], v. 585, p. 609-629, March 2022. Disponível em: https://doi.org/10.1016/j.ins.2021.11.036. Acesso em 26 dez. 2022.

LI, X.; ZHU, L.; ZHANG, C.; YANG, H.; WANG, H.; ZHANG, J.. Failure prediction for temporal dependency of hard drives. *In*: INTERNATIONAL WORKSHOP ON COMPUTER SCIENCE AND ENGINEERING (WCSE 2021), 11., 2021, Shanghai. **Proceedings [...]**. Shanghai: East China Normal University, China Agricultural University, Science and Engineering Institute, June, 2021. p. 379-388. Disponível em:

http://www.wcse.org/index.php?m=content&c=index&a=show&catid=22&id=1014. Acesso em: 24 ago. 2022.

MARTINS, M. E.G. Coeficiente de determinação. **Rev. Ciência Elem.**, [*s. l.*], v. 6, n. 01, p. 01, mar. 2018. Disponível em: http://doi.org/10.24927/rce2018.024. Acesso em: 25 dez. 2022.

McKEAN, J. W.; SIEVERS, G. L.. Coefficients of determination for least absolute deviation analysis. **Statistics & Probability Letters,** [*s. l.*], v. 5, issue1, p. 49-54 January 1987. Disponível em: https://doi.org/10.1016/0167-7152(87)90026-5. Acesso em: 27 dez. 2022.

NARAYANAN, I.; WANG, D.; JEON, M.; SHARMA, B.; CAULFIELD, L.; SIVASUBRAMANIAM, A.; CUTLER, B.; LIU, J.; KHESSIB, B.; VAID, K.. SSD failures in datacenters: What? When? and Why? **SYSTOR '16**: Proceedings of the 9th ACM International on Systems and Storage Conference, [*s. l.*], article n. 7, p. 1-11, June 2016. Disponível em: https://doi.org/10.1145/2928275.2928278. Acesso em 25 ago. 2022.

SANTOS, L. A. dos; MAZIERO, C. A.; BONA, L. C. E. de. Avaliação de caches em dispositivos de armazenamento secundário com SSDs. *In*: BRAZILIAN SYMPOSIUM ON COMPUTING SYSTEMS ENGINEERING (SBESC), 5., 2015, Foz do Iguaçu. **Anais eletrônicos** [...]. Foz do Iguaçu: UNIOESTE, nov. 2015. Disponível em: https://sbesc.lisha.ufsc.br/sbesc2015/proceedings/147458.pdf. Acesso em: 24 ago. 2022.

SHEN, J.; WAN, J.; LIM, S-J.; YU, L.. Random-forest-based failure prediction for hard disk drives. **International Journal of Distributed Sensor Networks**, [*s. l.*], v. 14, n. 11, p. 1-15, 2018. Disponível em: https://doi.org/10.1177/1550147718806480. Acesso em: 22 ago. 2022.

SHEN, J.; REN, Y.; WAN, J.; LAN, Y.. Hard disk drive failure prediction for mobile edge computing based on an LSTM recurrent neural network. **Mobile Information Systems**, [*s. l.*], v. 2021, article ID 8878364, p. 1-12, 2021. Disponível em: https://doi.org/10.1155/2021/8878364. Acesso em: 23 ago. 2022.